Reducing LLM Hallucinations!

Pavan Belagatti · Follow

Published in Level Up Coding · 8 min read · 4 days ago

13

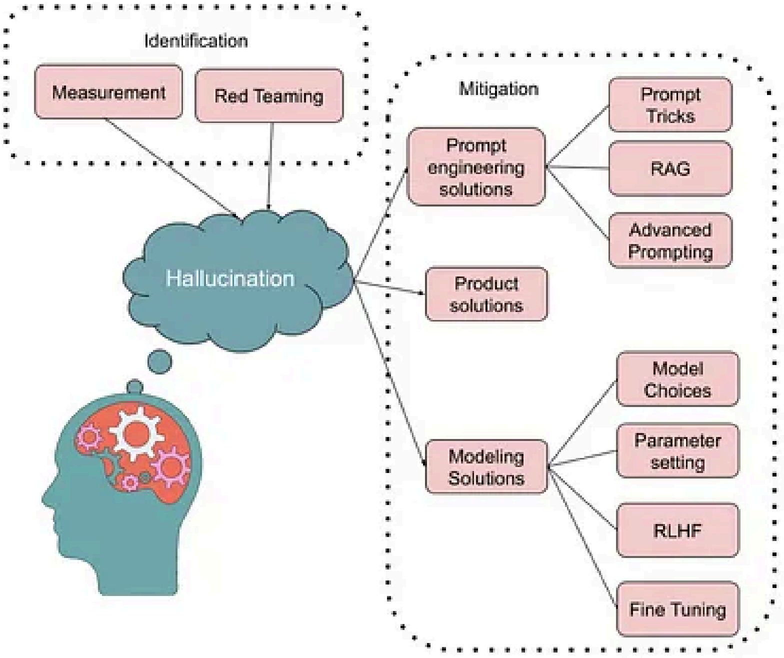


Image credits: Xavier(Xavi) Amatriain

Large language models (LLMs) have gained great momentum due to their inherent advantages for organizations building gen AI applications. While gen AI continues to impress as a shiny new area, it continues to evolve daily with new models in the market. While ChatGPT, LLama, Mistral and more get all the attention, we also need to consider how effective these LLMs perform for a specific task. And occasionally, LLMs produce incorrect responses that are presented confidently. This is one of the major challenges in using these LLMs in building applications — for various reasons, they hallucinate. The good news? There is a way to reduce these incidents.

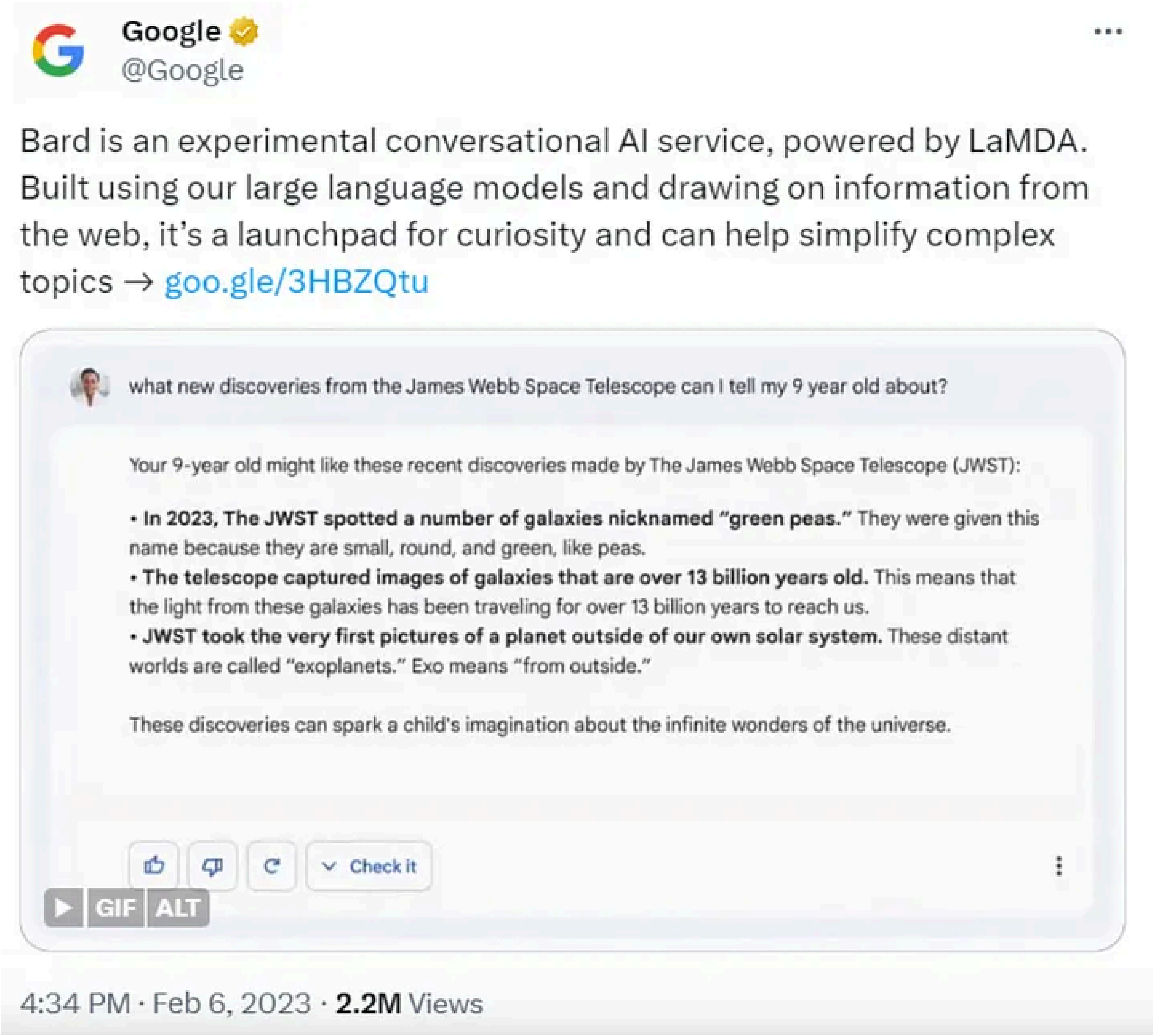
Let's dive deeper into LLM hallucinations, and understand how you can efficiently reduce this phenomenon with certain techniques.

What is hallucination in Large Language Models?

Hallucination is a phenomenon where an AI model produces text that, while often grammatically correct and plausible, is not rooted in the given input — or may even be factually incorrect.

LLMs have completely revolutionized the way we interact with machines by making them more wise and efficient to complete different custom tasks. But, it is the inherent property of any LLM to sometimes generate incorrect answers that are correct — including ones with biases or factually incorrect responses. This happens due to several factors including the data they are trained on, low quality input, poor prompts, etc.

You might be wondering how bad the impact of these hallucinations can be. The following screenshot shows Google's Bard (now Gemini) generating hallucinated answers and responses. When one user asked about a query, Bard falsely stated something regarding the James Webb Space Telescope.



Google
@Google

Bard is an experimental conversational AI service, powered by LaMDA. Built using our large language models and drawing on information from the web, it's a launchpad for curiosity and can help simplify complex topics → goo.gle/3HBZQtu

what new discoveries from the James Webb Space Telescope can I tell my 9 year old about?

Your 9-year old might like these recent discoveries made by The James Webb Space Telescope (JWST):

- **In 2023, The JWST spotted a number of galaxies nicknamed "green peas."** They were given this name because they are small, round, and green, like peas.
- **The telescope captured images of galaxies that are over 13 billion years old.** This means that the light from these galaxies has been traveling for over 13 billion years to reach us.
- **JWST took the very first pictures of a planet outside of our own solar system.** These distant worlds are called "exoplanets." Exo means "from outside."

These discoveries can spark a child's imagination about the infinite wonders of the universe.

4:34 PM · Feb 6, 2023 · **2.2M** Views

Image source: X

This hallucination from Bard cost Google a nearly $100B drop in valuation.

Why do hallucinations occur in LLMs?

As previously discussed, language models can hallucinate and produce outputs that include made-up or incorrect responses. These errors demonstrate the limitations of AI, highlighting the importance of human oversight and cross-checking against reliable sources for verification. But assigning a human to verify each response is not feasible or scalable. We will talk about hallucination mitigation strategies in a minute, but first let's see why hallucinations occur in LLMs:

- **Insufficient training data.** A model that hasn't encountered diverse data during training might not establish accurate correlations between inputs and appropriate outputs, leading to hallucinated content.
- **Inadequate supervision.** Without proper guidance, a model might rely too heavily on its internal logic, leading to outputs that appear to hallucinate.
- **Model overfitting.** Overfitting to training data can cause a model to produce outputs that mirror the training set, but are misaligned with new or different inputs.
- **Knowledge cutoff.** LLMs like ChatGPT have a knowledge cutoff date and thus, are unaware of any information past that date. They may unknowingly respond to your question with out-of-date information that is no longer relevant.

Types of LLM hallucinations

We can divide these hallucination types into three main categories:

- **Factual inaccuracies.** This type of hallucination occurs when a language model presents information that is not true or correct, but is framed as if factual. This includes dates, events, statistics or statements that are verifiably false. It can happen due to various reasons including misinterpretation of input data, low quality data and training methodologies, reliance on outdated or incorrect sources or the blending of information from different contexts that leads to an inaccurate output.

- **Generated quotations or sources.** This occurs when a language model fabricates quotes or citations. It might generate a statement and incorrectly attribute it to a real person, or create a fictitious source that does not exist at all. This is problematic because it leads to misinformation, falsely attributed statements and confusion.
- **Logical inconsistencies.** This includes generating responses that are internally inconsistent or logically flawed. After generating a response for a user query, the LLM can contradict itself in further responses. It occurs when a model makes a series of statements that, when taken together, are incoherent or conflicting — challenging the credibility of the model's outputs and confusing users who rely on its consistency.

In all of these cases, the language model is not intentionally misleading but is exhibiting its own limitations from various reasons that might include its training data, quality of data, knowledge cut-off date, poor fine-tuning, etc. And as a result, AI researchers are now coming up with various frameworks and tools to mitigate these hallucinations.

LLM hallucination mitigation strategies

Researchers are developing various approaches to make sure the response generated by LLMs is accurate. Some strategies need human intervention like reinforcement learning through human feedback (RLHF); others need fresh and custom data to train the model, known as fine-tuning. While Retrieval Augmented Generation (RAG) can help reduce hallucinations generated by LLMs, let's look at some more in-depth approaches we can use.

Along with RAG, we can divide these strategies into two parts: Pre-generation and post-generation strategies.

Pre-generation strategies prevent AI from generating incorrect or misleading information in the first place. These include:

Prompting

- **Chain-of-Verification (CoVe).** This involves self verification of responses by models. Multiple stages of verification makes it more efficient.
- **Optimization by PROmpting (OPRO).** This is where LLMs tend to optimize their own prompts, correcting the prompt inputs.
- **System 2 Attention (S2A).** This approach improves LLM reasoning. An instruction-tuned LLM is used here to identify, analyze and extract the most relevant parts of the input context, mitigating the influence of unnecessary information.
- **EmotionPrompt.** This technique uses emotional cues through prompts to LLMs so they can have more context and sentiment.
- **Step-Back Prompting.** This is an approach used to improve LLM reasoning and problem-solving skills.
- **Rephrase and Respond (RaR)**. This technique allows LLMs to rephrase and expand the questions/prompts posed by humans, helping LLMs gain insightful context.

Retrieval Augmented Generation (RAG)

- **Self-RAG.** Self-RAG empowers LLMs to dynamically fetch relevant passages until the entire context is captured, all within the specified window.
- **Active-RAG.** Active-RAG improves passive RAG by fine-tuning the retriever based on feedback from the generator during multiple interactions.

- **Multimodal RAG.** Multimodal RAG gives a deeper understanding of context by augmenting text data with images and other media, enabling more accurate and relevant responses.

Post-generation strategies deal with verifying and correcting the AI's outputs after they have been generated. These include:

- **Fact checking.**: Implementing human-in-the-loop (HITL) and knowledge bases to verify the accuracy of the information provided by the LLMs.
- **Preference alignment.** Using human feedback mechanisms (RLHF) to align the LLM's outputs with human values and preferences.

These strategies are aimed at enhancing the reliability of AI systems, improving the quality of their outputs and ensuring they are aligned with human values and factual accuracy.

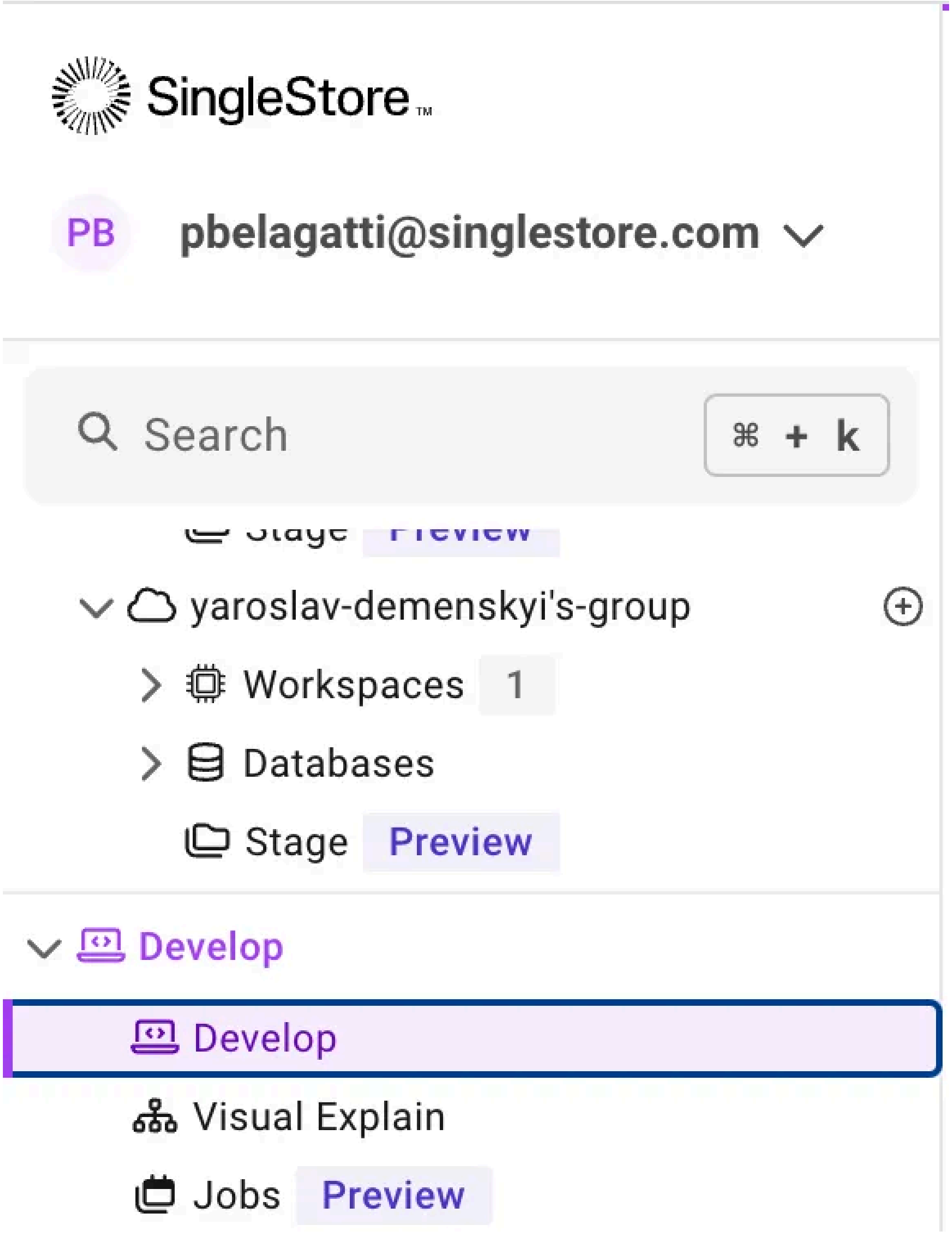
Mitigating hallucinations in your LLM apps with SingleStore

In this tutorial, we will take a publicly available news dataset, storing it in our SingleStore database and retrieving the information through hybrid search, a classic RAG approach. Through this method, we aim to mitigate instances of hallucination, ensuring precise and relevant results.

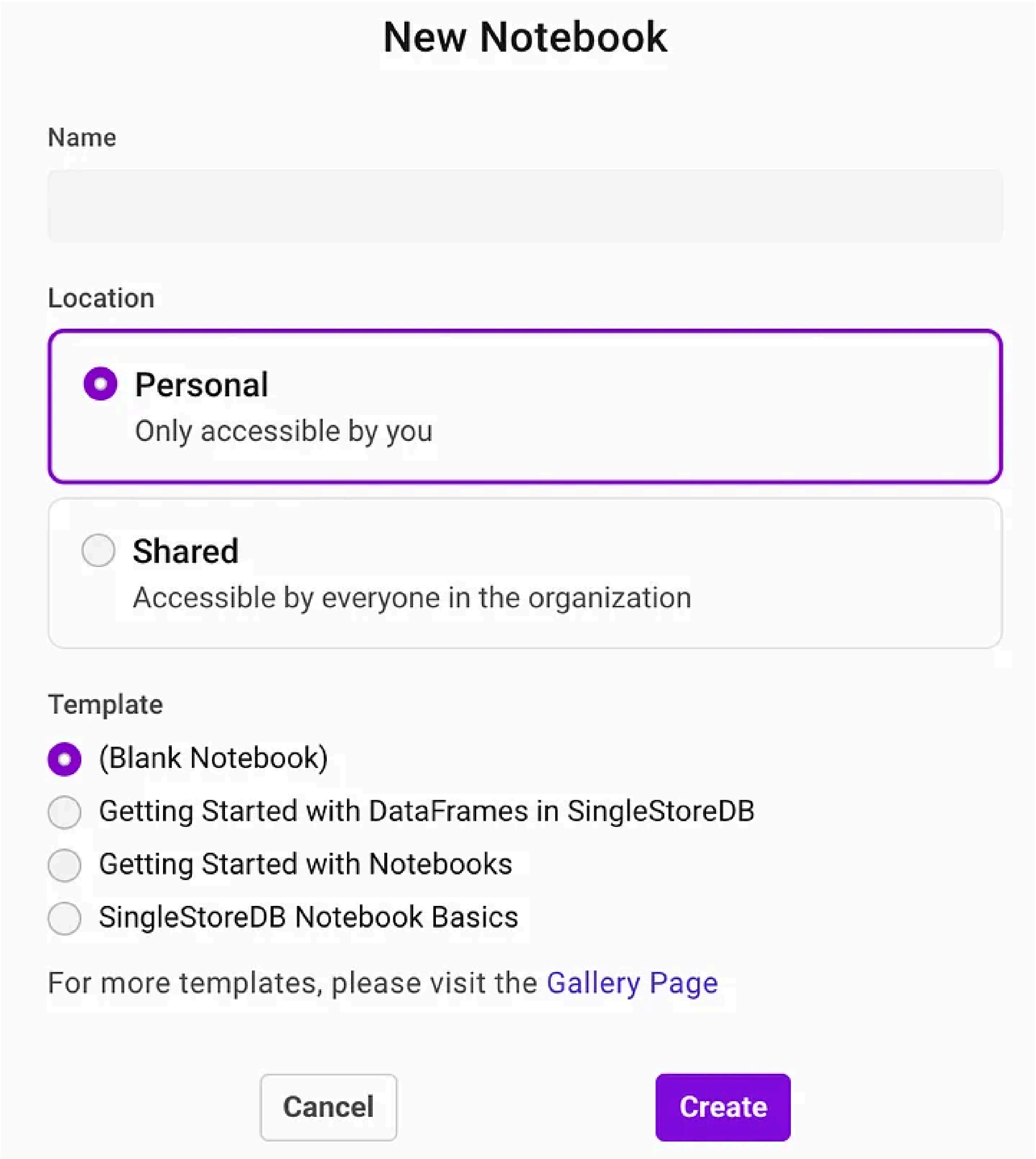
We will use SingleStore's Notebook feature to complete this tutorial.

If you haven't already, activate your free SingleStore trial to use Notebooks and create a database to store our embeddings (vector) data.

Once you sign up, you land on the SingleStore dashboard. Click on 'Develop' to create a Notebook.



Create a new Notebook, naming it whatever you want.



New Notebook

Name

Location

● Personal
Only accessible by you

○ Shared
Accessible by everyone in the organization

Template

● (Blank Notebook)
○ Getting Started with DataFrames in SingleStoreDB
○ Getting Started with Notebooks
○ SingleStoreDB Notebook Basics

For more templates, please visit the Gallery Page

Start with installing and importing the required libraries

```
!pip3 install wget --quiet
!pip3 install openai==1.3.3 --quiet
```

```
!pip3 install sentence-transformers --quiet
```

```
import json
import os
import pandas as pd
import wget
```

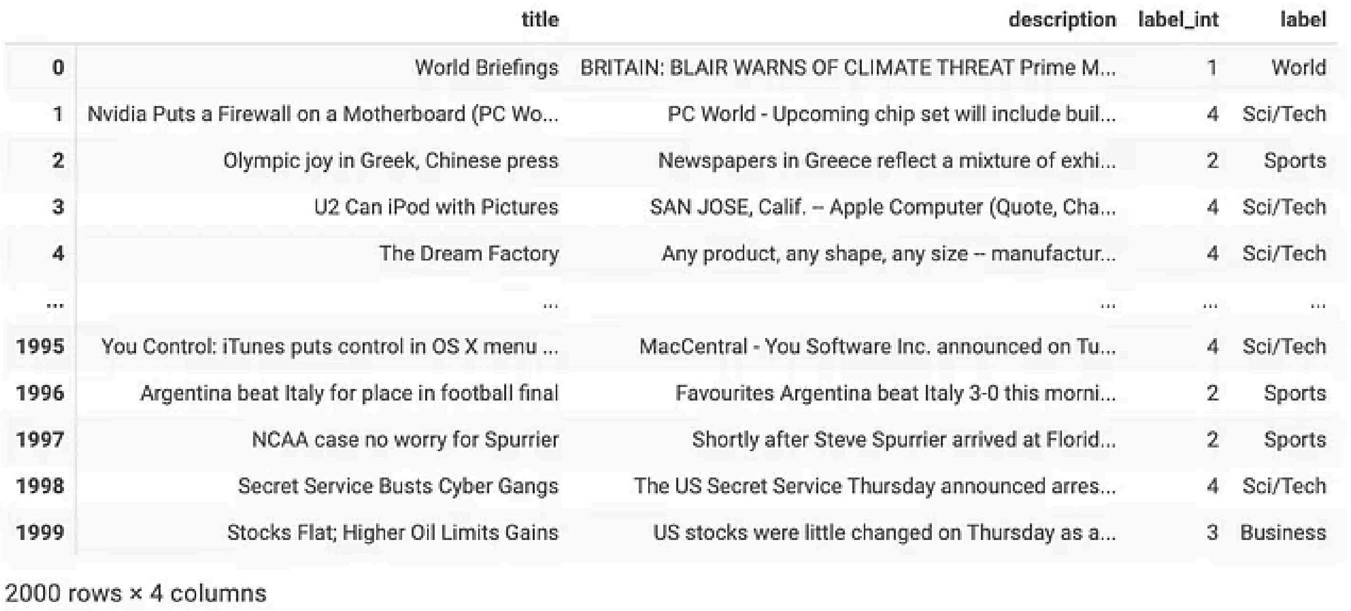
Download the model

```
from sentence_transformers import SentenceTransformer
model = SentenceTransformer('flax-sentence-embeddings/all_datasets_v3_mpnet-base
```

Import data from the csv file (AG News is a subdataset of AG's corpus of news articles)

```
cvs_file_path = 'https://raw.githubusercontent.com/openai/openai-cookbook/main/e
file_path = 'AG_news_samples.csv'

if not os.path.exists(file_path):
    wget.download(cvs_file_path, file_path)
    print('File downloaded successfully.')
else:
    print('File already exists in the local file system.')

df = pd.read_csv('AG_news_samples.csv')
df
```

This is what you should see after running the preceding code to read the contents of the csv file.



	title	description	label_int	label
0	World Briefings	BRITAIN: BLAIR WARNS OF CLIMATE THREAT Prime M...	1	World
1	Nvidia Puts a Firewall on a Motherboard (PC Wo...	PC World - Upcoming chip set will include buil...	4	Sci/Tech
2	Olympic joy in Greek, Chinese press	Newspapers in Greece reflect a mixture of exhi...	2	Sports
3	U2 Can iPod with Pictures	SAN JOSE, Calif. -- Apple Computer (Quote, Cha...	4	Sci/Tech
4	The Dream Factory	Any product, any shape, any size -- manufactur...	4	Sci/Tech
...	...	...	...	...
1995	You Control: iTunes puts control in OS X menu ...	MacCentral - You Software Inc. announced on Tu...	4	Sci/Tech
1996	Argentina beat Italy for place in football final	Favourites Argentina beat Italy 3-0 this morni...	2	Sports
1997	NCAA case no worry for Spurrier	Shortly after Steve Spurrier arrived at Florid...	2	Sports
1998	Secret Service Busts Cyber Gangs	The US Secret Service Thursday announced arres...	4	Sci/Tech
1999	Stocks Flat; Higher Oil Limits Gains	US stocks were little changed on Thursday as a...	3	Business

2000 rows × 4 columns

You can see the data here

```
data = df.to_dict(orient='records')
data[0]
```

The next step is setting up the database to store our data

```
shared_tier_check = %sql show variables like 'is_shared_tier'
if not shared_tier_check or shared_tier_check[0][1] == 'OFF':
    %sql DROP DATABASE IF EXISTS news;
    %sql CREATE DATABASE news;

%%sql
DROP TABLE IF EXISTS news_articles;
CREATE TABLE IF NOT EXISTS news_articles (
    title TEXT,
    description TEXT,
    genre TEXT,
    embedding BLOB,
    FULLTEXT(title, description)
);
```

Get embeddings for every row based on the description column

```
descriptions = [row['description'] for row in data]
all_embeddings = model.encode(descriptions)
all_embeddings.shape
```

Merge embedding values into data rows

```
for row, embedding in zip(data, all_embeddings):
    row['embedding'] = embedding
```

Here is an example of one row of the combined data

```
data[0]
```

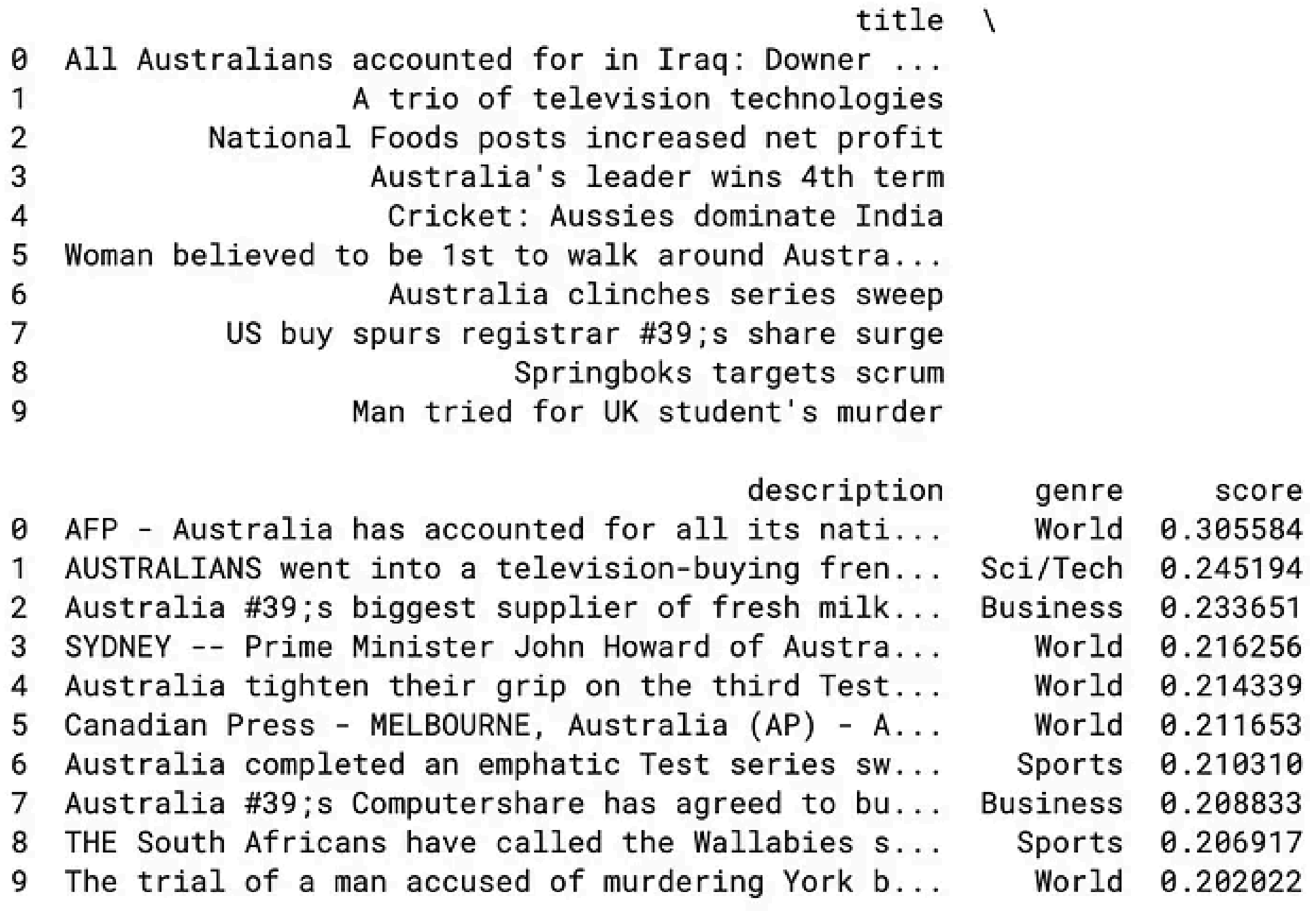
Now, let’s populate the database with our data

```
%sql TRUNCATE TABLE news_articles;

import sqlalchemy as sa
from singlestoredb import create_engine

# Use create_table from singlestoredb since it uses the notebook connection URL
conn = create_engine().connect()

statement = sa.text('''
    INSERT INTO news_articles (
        title,
        description,
        genre,
        embedding
    )
    VALUES (
        :title,
        :description,
        :label,
        :embedding
    )
''')

conn.execute(statement, data)
```

Let’s run semantic search, and get scores for the search term ‘Aussie’

```
search_query = 'Aussie'
search_embedding = model.encode(search_query)

query_statement = sa.text('''
    SELECT
        title,
        description,
        genre,
        DOT_PRODUCT(embedding, :embedding) AS score
    FROM news_articles
    ORDER BY score DESC
    LIMIT 10
''')
```

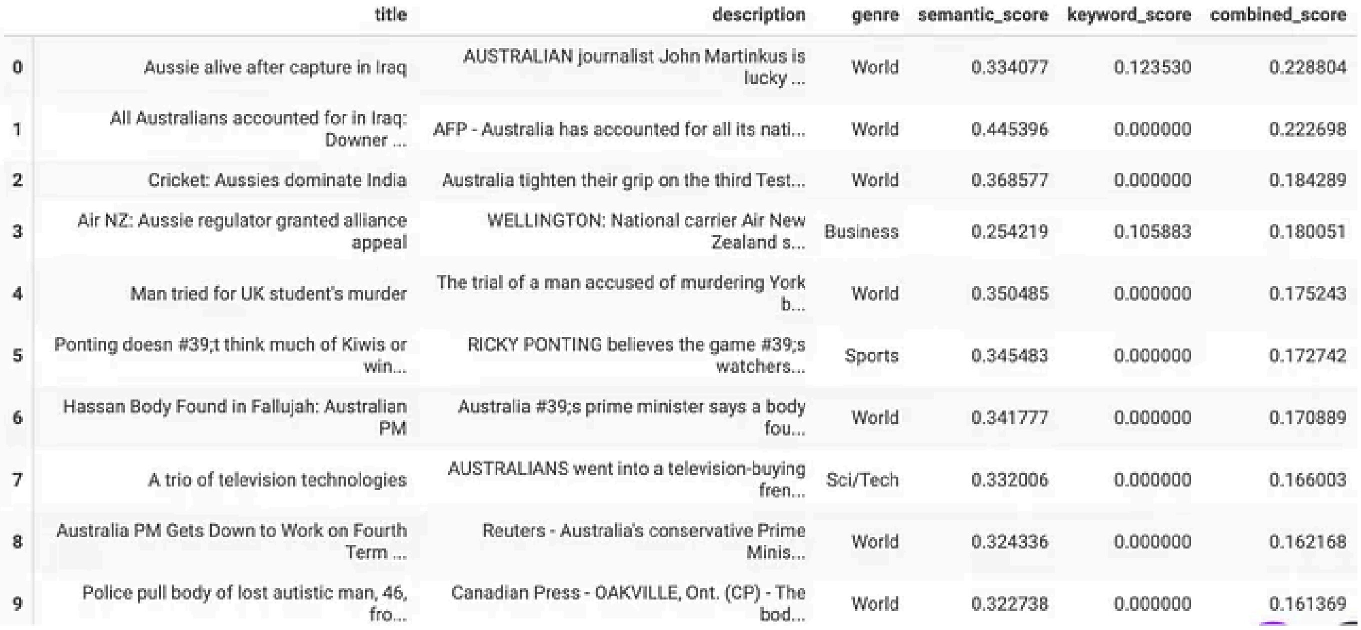
```
    # Execute the SQL statement.
    results = pd.DataFrame(conn.execute(query_statement, dict(embedding=search_embed
    print(results)
```



```
                                               title  \
0  All Australians accounted for in Iraq: Downer ...
1                A trio of television technologies
2            National Foods posts increased net profit
3                  Australia's leader wins 4th term
4                 Cricket: Aussies dominate India
5  Woman believed to be 1st to walk around Austra...
6                   Australia clinches series sweep
7             US buy spurs registrar #39;s share surge
8                          Springboks targets scrum
9                Man tried for UK student's murder

                                         description     genre     score
0  AFP - Australia has accounted for all its nati...     World  0.305584
1  AUSTRALIANS went into a television-buying fren...  Sci/Tech  0.245194
2  Australia #39;s biggest supplier of fresh milk...  Business  0.233651
3  SYDNEY -- Prime Minister John Howard of Austra...     World  0.216256
4  Australia tighten their grip on the third Test...     World  0.214339
5  Canadian Press - MELBOURNE, Australia (AP) - A...     World  0.211653
6  Australia completed an emphatic Test series sw...    Sports  0.210310
7  Australia #39;s Computershare has agreed to bu...  Business  0.208833
8  THE South Africans have called the Wallabies s...    Sports  0.206917
9  The trial of a man accused of murdering York b...     World  0.202022
```

Now, let’s run a hybrid search to find articles about Aussie captures.

```
    hyb_query = 'Articles about Aussie captures'
    hyb_embedding = model.encode(hyb_query)

    # Create the SQL statement.
    hyb_statement = sa.text('''
        SELECT
            title,
            description,
            genre,
            DOT_PRODUCT(embedding, :embedding) AS semantic_score,
            MATCH(title, description) AGAINST (:query) AS keyword_score,
            (semantic_score + keyword_score) / 2 AS combined_score
        FROM news_articles
        ORDER BY combined_score DESC
        LIMIT 10
    ''')

    # Execute the SQL statement.
    hyb_results = pd.DataFrame(conn.execute(hyb_statement, dict(embedding=hyb_embedd
    hyb_results
```



	title	description	genre	semantic_score	keyword_score	combined_score
0	Aussie alive after capture in Iraq	AUSTRALIAN journalist John Martinkus is lucky ...	World	0.334077	0.123530	0.228804
1	All Australians accounted for in Iraq: Downer ...	AFP - Australia has accounted for all its nati...	World	0.445396	0.000000	0.222698
2	Cricket: Aussies dominate India	Australia tighten their grip on the third Test...	World	0.368577	0.000000	0.184289
3	Air NZ: Aussie regulator granted alliance appeal	WELLINGTON: National carrier Air New Zealand s...	Business	0.254219	0.105883	0.180051
4	Man tried for UK student's murder	The trial of a man accused of murdering York b...	World	0.350485	0.000000	0.175243
5	Ponting doesn #39;t think much of Kiwis or win...	RICKY PONTING believes the game #39;s watchers...	Sports	0.345483	0.000000	0.172742
6	Hassan Body Found in Fallujah: Australian PM	Australia #39;s prime minister says a body fou...	World	0.341777	0.000000	0.170889
7	A trio of television technologies	AUSTRALIANS went into a television-buying fren...	Sci/Tech	0.332006	0.000000	0.166003
8	Australia PM Gets Down to Work on Fourth Term ...	Reuters - Australia's conservative Prime Minis...	World	0.324336	0.000000	0.162168
9	Police pull body of lost autistic man, 46, fro...	Canadian Press - OAKVILLE, Ont. (CP) - The bod...	World	0.322738	0.000000	0.161369

By leveraging targeted mitigation tools and strategies like RAG, we can significantly diminish instances of hallucinations. Reducing hallucinations enhances the reliability and accuracy of LLM-powered applications — and also propels us closer to harnessing the full capabilities of generative AI. The journey toward perfecting these models is ongoing, but with constant effort and innovative solutions like SingleStore, we can maximize their benefit for a wide array of applications.

If you are interested in understanding more about the concept of LLM hallucinations, check out on-demand webinar.

Activate your SingleStore free trial to try the above tutorial.

Large Language Models

Retrieval Augmented

Llm

Llm Applications

Vector Database

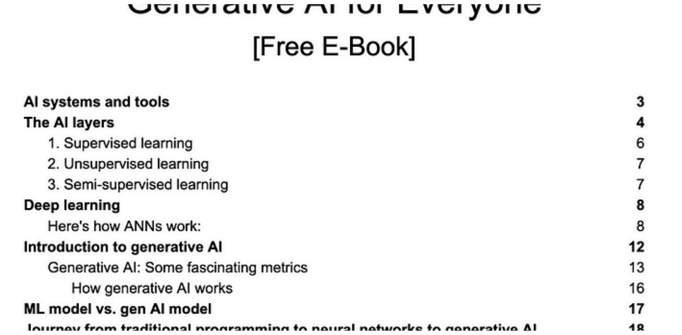
Written by Pavan Belagatti

Follow

3.8K Followers · Writer for Level Up Coding

Developer Evangelist | AI/ML| DevOps | Data Science! Currently working at SingleStore as a Developer Evangelist.

More from Pavan Belagatti and Level Up Coding



Generative AI for Everyone
[Free E-Book]

AI systems and tools	**3**
The AI layers	**4**
1. Supervised learning	6
2. Unsupervised learning	7
3. Semi-supervised learning	7
Deep learning	**8**
Here's how ANNs work:	8
Introduction to generative AI	**12**
Generative AI: Some fascinating metrics	13
How generative AI works	16
ML model vs. gen AI model	**17**
Journey from traditional programming to neural networks to generative AI	**18**

Pavan Belagatti in Level Up Coding

Generative AI for Everyone!

Contents of this E-Book

4 min read · Apr 11, 2024

68

Daniel Craciun in Level Up Coding

Stop Using TypeScript Interfaces

Why You Should Use Types Instead

★ · 4 min read · Apr 16, 2024

533 21

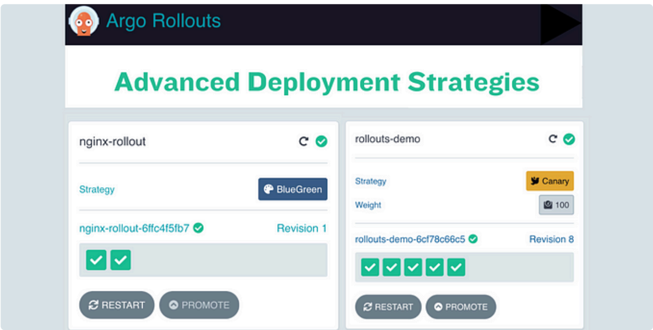
Liu Zuo Lin in Level Up Coding

5 SQL Things I Should Have Known Earlier But Somehow Didn’t

This might help you in SQL interviews

★ · 4 min read · Mar 16, 2024

1.5K 10



Pavan Belagatti in Data And Beyond

Argo Rollouts: Advanced Strategies for Smooth...

In the fast-paced world of modern software development, Kubernetes has emerged as a...

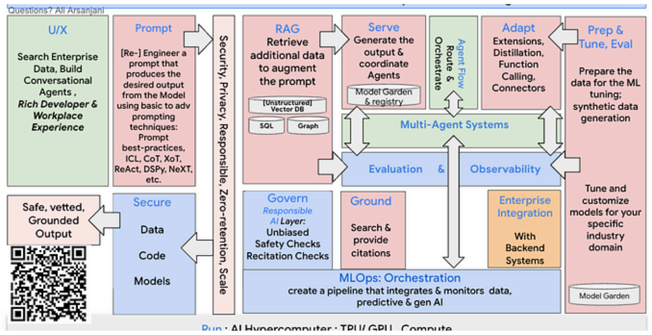
6 min read · Aug 2, 2023

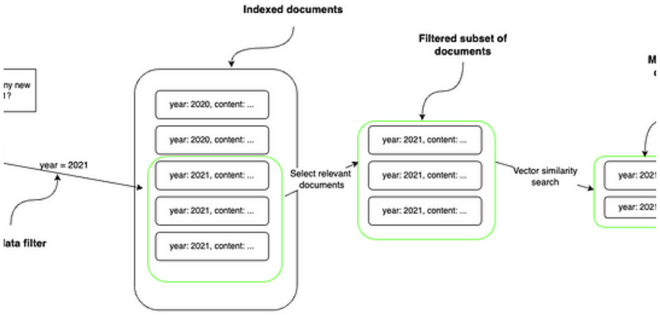
53

See all from Pavan Belagatti

See all from Level Up Coding

Recommended from Medium


U/X
Search Enterprise Data, Build Conversational Agents , Rich Developer & Workplace Experience
Prompt
[Re-] Engineer a prompt that produces the desired output from the Model using basic to adv prompting techniques: Prompt best-practices, ICL, CoT, XoT, ReAct, DSPy, NeXT, etc.
Security, Privacy, Responsible, Zero-retention, Scale
RAG
Retrieve additional data to augment the prompt
[Unstructured] Vector DB
SQL
Graph
Serve
Generate the output & coordinate Agents
Model Garden & registry
Route & Orchestrate
Agent Flow
Adapt
Extensions, Distillation, Function Calling, Connectors
Prep & Tune, Eval
Prepare the data for the ML tuning; synthetic data generation
Multi-Agent Systems
Evaluation & Observability
Safe, vetted, Grounded Output
Secure
Data
Code
Models
Govern
Responsible AI Layer:
Unbiased Safety Checks Recitation Checks
Ground
Search & provide citations
Enterprise Integration
With Backend Systems
Tune and customize models for your specific industry domain
Model Garden
MLOps: Orchestration
create a pipeline that integrates & monitors data, predictive & gen AI



Ali Arsanjani

The GenAI Reference Architecture

In this article we are providing the major architectural building blocks and blueprint f...

26 min read · 5 days ago

35 1

Tomaz Bratanic in Neo4j Developer Blog

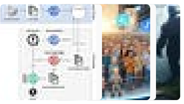
Graph-based Metadata Filtering for Improving Vector Search in RAG...

Optimizing vector retrieval with advanced graph-based metadata techniques using...

11 min read · 5 days ago

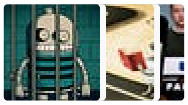
235

Lists



Natural Language Processing

1424 stories · 921 saves

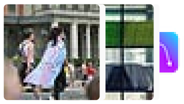


AI Regulation

6 stories · 434 saves

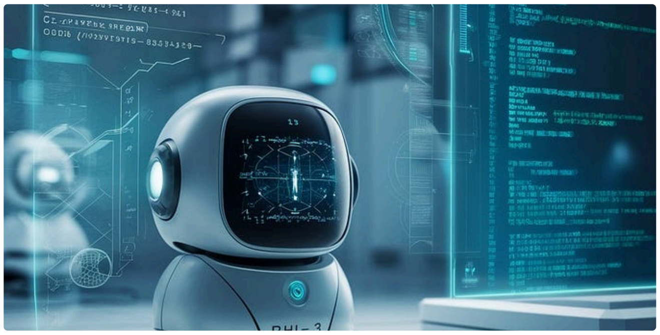
ChatGPT prompts

47 stories · 1508 saves



Generative AI Recommended Reading

52 stories · 993 saves



Jesus Rodriguez in Towards AI

Some Technical Notes About Phi-3: Microsoft’s Marquee Small…

The model ius able to outperform much larger alternatives and now run locally on mobile…

4 min read · 5 days ago

6

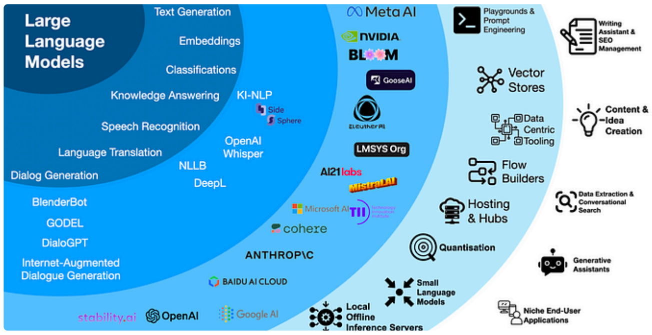
Matthias Dittrich in UX Planet

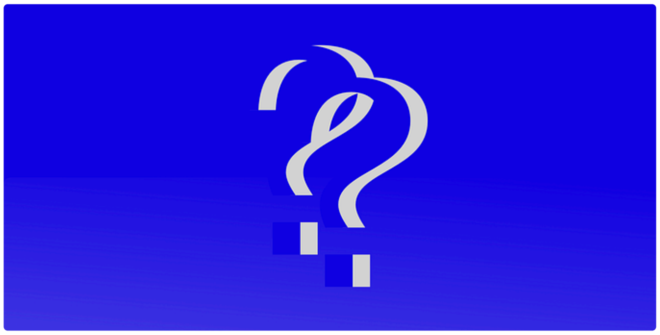
Stop designing chat-based AI tools.

It is time to evolve AI tools beyond prompt-based interfaces and consider new mental…

5 min read · Apr 21, 2024

921 19





Cobus Greyling

The Large Language Model Landscape—Version 5

In the recent past I have been observing and describing current LLM-related technologie…

5 min read · Apr 25, 2024

62 2

Elaine Lu in Towards Data Science

Why Do AI Projects Fail?

85% AI projects fail, 6 reasons why

9 min read · 4 days ago

364 9

See more recommendations